Safe Adaptation in Uncertain Constrained Markov Decision Processes

RESEARCH OUTLINE

Yarden As

Jun. 16th 2023

Yarden As

PhD student @ ETH AI Center & Learning & Adaptive Systems.
Working on safe reinforcement learning.

Passionate about making reinforcement learning work in the real world.

KGLW fan / (former) rock clibmer / left-handed.

1991
Born, Haifa, Israel

2010
IDF (Rimon)

2014
Mechanical Engineering, Technion

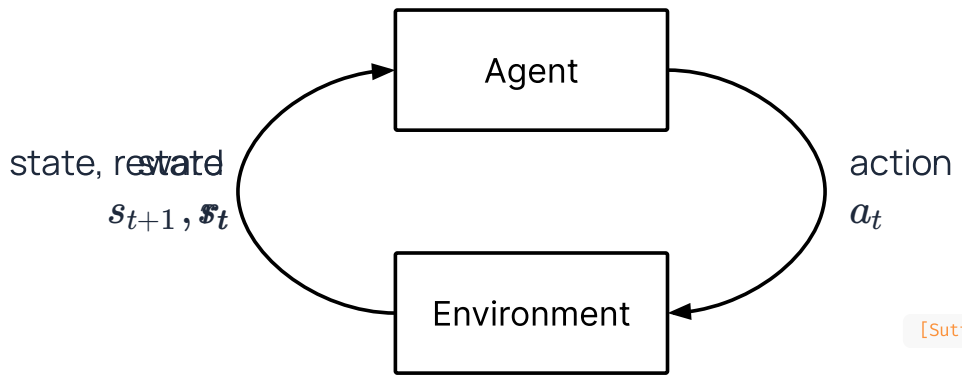
2017
Robotics Engineer, Tel-Aviv

2019
Robotics, Systems and Control, ETH Zürich

2022
Doctoral Fellow, ETH Zürich

Back to Basics

Reinforcement learning: a data-driven approach for sequential decision-making



[Sutton & Barto (1992)]

Recap

Interacting with the environment is modeled as a Markov Decision Process

- State space (discrete): $s_t \in \mathcal{S}$
- Action space (discrete): $a_t \in \mathcal{A}$
- Transition function: $s_{t+1} \sim P(\cdot|s_t, a_t), P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$
- Reward function: $r_t = R(s_t, a_t), R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Policy: $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$
- Trajectory $\tau = \{s_0, a_0, a_1, \ldots, a_{T-1}, s_T\}, \tau \sim p(\tau) = p(s_0) \prod_{t=1}^{T} \pi(a_t|s_t) P(s_{t+1}|s_t, a_t)$

Goal: find a policy $\pi(a_t|s_t)$

$$\pi \in \arg\max_{\pi} \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} R(s_t, a_t) \right]$$

Is it a perfect model?

Some Open Challenges

What if $P(s'|s,a)$ or $R(s_t,a_t)$ vary between trials?

Is the reward enough to induce *safe* behavior?

Real-World Examples

Where do these challeges emerge?

Agriculture & Sustainability

- Different weather regimes and soil compositions.
- Over-fertilization accelerates greenhouse effects. [Erisman et al., (2008)]

Medical Applications

- Patients react differently to different treatments.
- Clinical setting requires safety. [Pace et al., (2022)]

Robotics

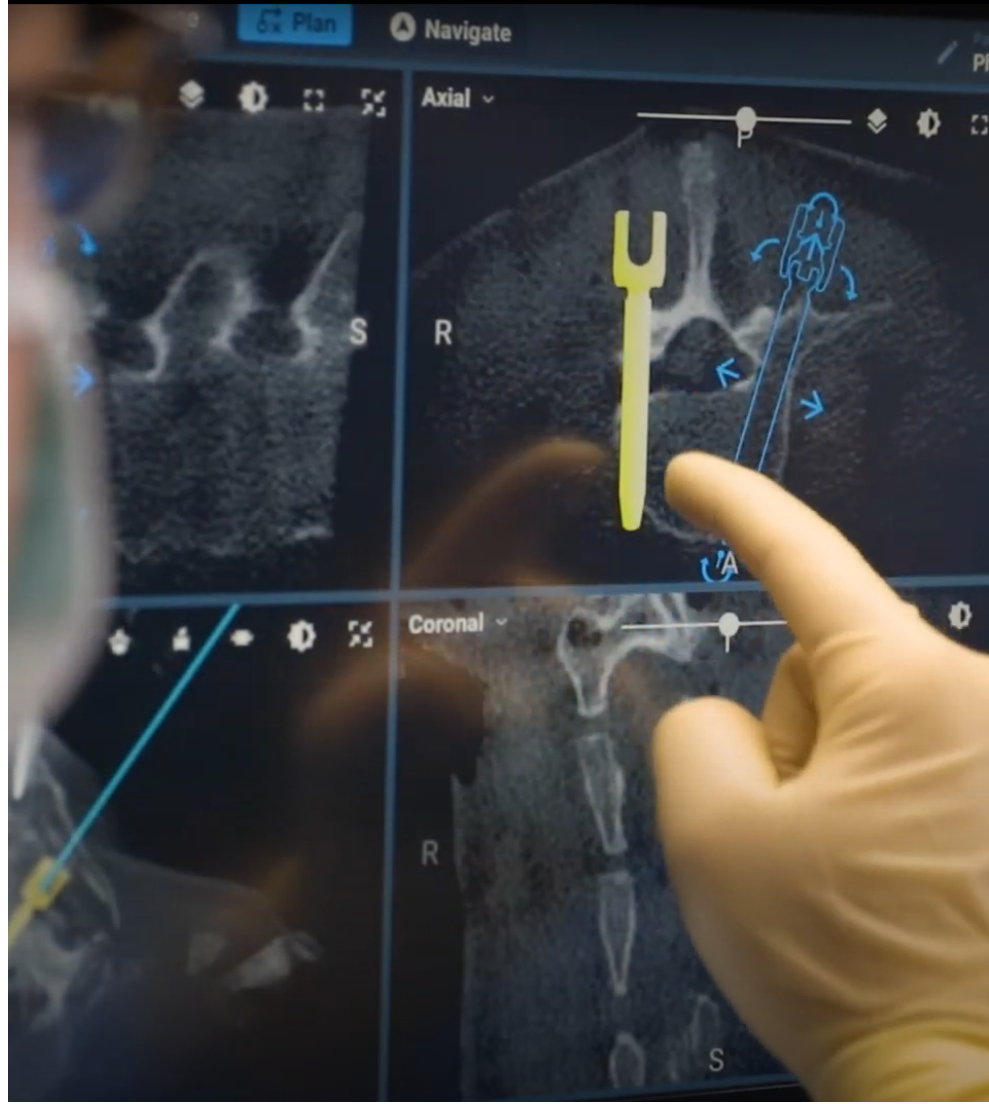
- Objectives and dynamics may vary. [Thrun et al., (2005)]
- Robots should not harm themselves or their environment.

Automatic Spine Surgery

Key challenges remain

- *Orthopedic variation* of patients create similar, but different, planning problems.
- Clinical setting requires high-precision and safety.

$\Longrightarrow$ *safe* and *adaptive* planning algorithms.



Constrained Markov Decision Processes (CMDP)

A short introduction

Idea: cost signal c_t together with the reward r_t.

- Cost function: $c_t = C(s_t, a_t), C : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Goal: find a policy $\pi(a_t | s_t)$ that solves the *constrained* problem

$$\max_{\pi} \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} R(s_t, a_t) \right]$$

$$\text{s.t. } \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_{t=0}^{T} C(s_t, a_t) \right] \leq 0$$

Meta-Learning

A framework for data-efficient adaptation

Central concept: use data from multiple, related, tasks to learn informative priors for future tasks.

- Meta-train data: $\mathcal{D}_k = \{x_i, y_i\}_{i=1}^N$
- $k = 1 \dots K$ tasks
- $y_i = f_{\theta^k}(x_i) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma)$
- Meta-test data: $\tilde{\mathcal{D}} = \{x_j, f_{\tilde{\theta}}(x_j) + \epsilon\}_{j=1}^M$
- Learn prior $p(\tilde{\theta})$ with $\mathcal{D}_{1:K}$. Use it to infer $p(\tilde{\theta}|\tilde{\mathcal{D}})$ more efficiently. [Rothfuss et al. (2021)]

Strong foundation on safety
and meta-learning,
but not on their intersection.

Research Goal

Devise algorithms that address the key challenges of safe adaptation with the aim of making them applicable in the field of robotic spinal surgery.

Adaptation to Variable CMDPs

Via meta-reinforcement learning

- "Meta-environment": $\mathcal{M}_i \sim \mathcal{E}(\cdot)$
- CMDP (informally): $\mathcal{M}_i = (P_i(s'|s,a,), C_i(s,a), R_i(s,a))$

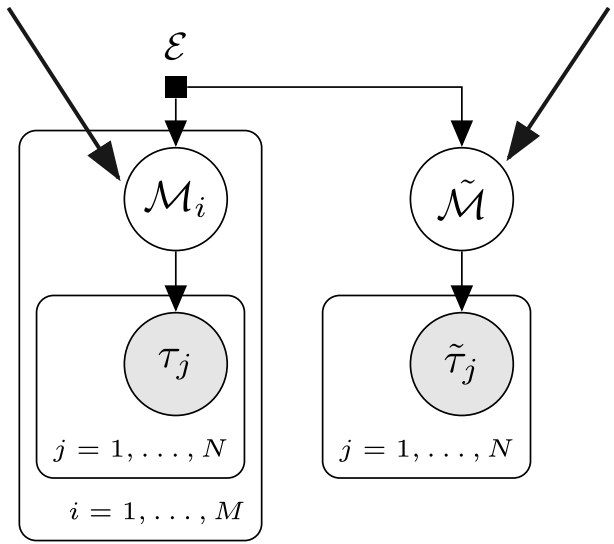
During training: interact with $\mathcal{M}_i, i = 1, \ldots, M$ CMDPs.

During testing: interact with $\tilde{\mathcal{M}}$

Goal: *adapt* to the problem induced by $\tilde{\mathcal{M}}$ and solve

$$\max_{\pi} \mathbb{E}_{\tilde{\tau} \sim p(\tilde{\tau})} \left[\sum_{t=0}^{T} R(s_t, a_t) \right]$$

$$\text{s.t. } \mathbb{E}_{\tilde{\tau} \sim p(\tilde{\tau})} \left[\sum_{t=0}^{T} C(s_t, a_t) \right] \leq 0$$



Solving CMDPs

(w/o transfer to new tasks)

Constrained Policy Optimization via Bayesian World Models [As et al. (2022)]

LAMBDA

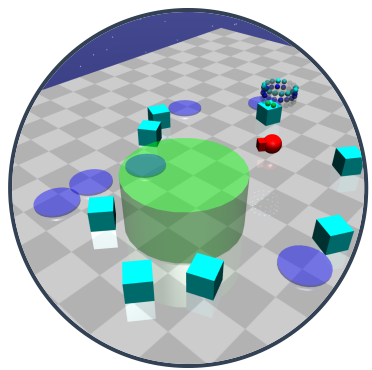
- Learn a Bayesian model of $P(s'|s,a)$
- Use it for policy optimization
- Solve constrained problem via Augmented Lagrangian [Bertsekas, Dimitri P. (1996)]

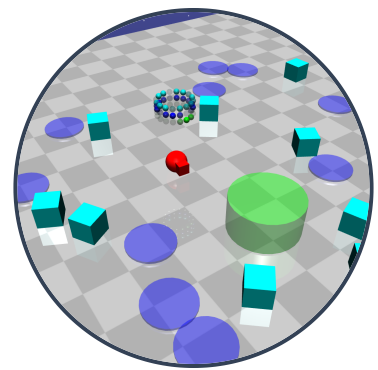
Solving CMDPs

(w transfer to new tasks)

Log Barriers for Safe Black-box Optimization with Application to Safe Reinforcement Learning [Usmanova et al. (2022)]

Challenge: given an initially safe policy,
how to transfer it to a new task while maintaining safety?





Summary

Challenge

SAFE ADAPTATION IS A COMMON, YET, AN OPEN PROBLEM.

END

Appendix

Spine Surgery & RL

What's the motivation to use RL?

Online planning and control from noisy observations problem.

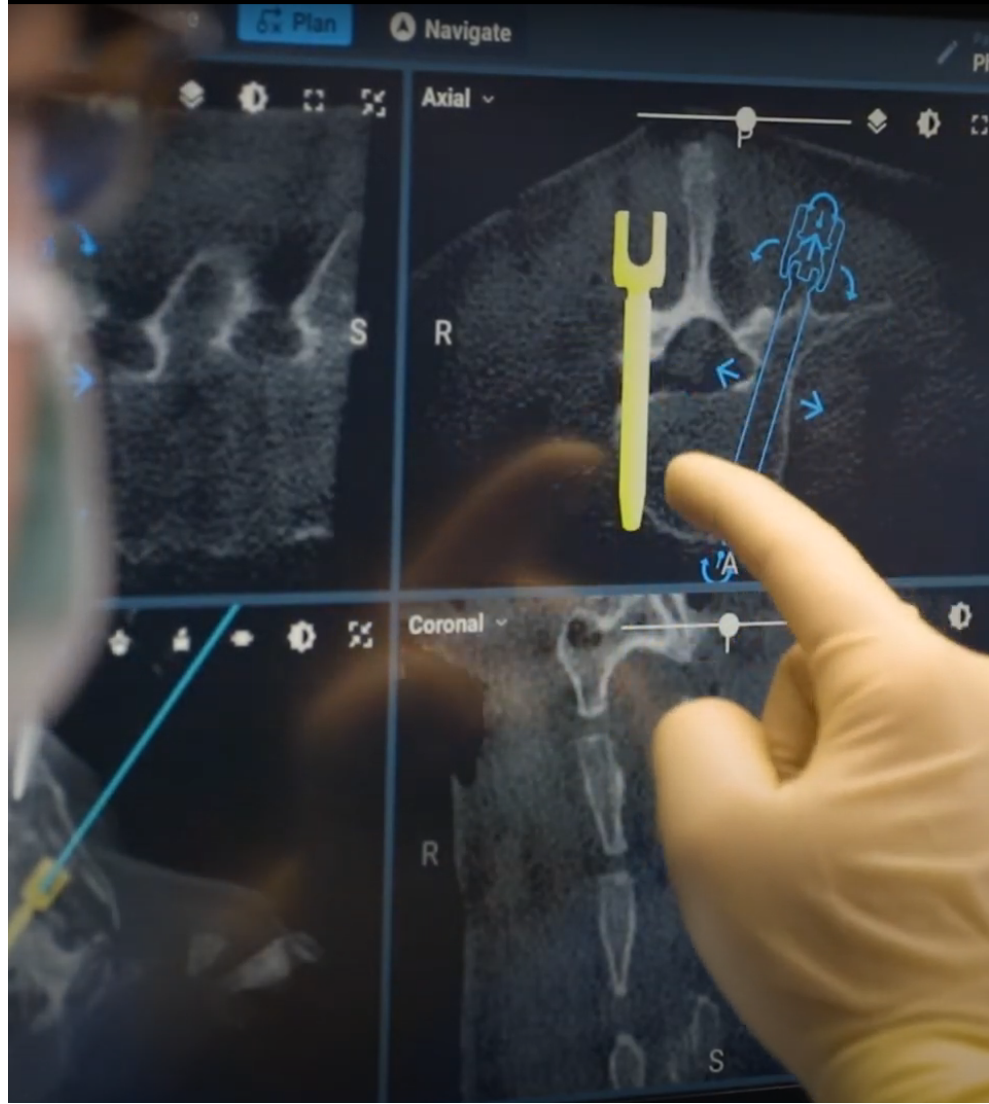
Limitations

- Invasive imaging (CT, X-Ray).
- Pre-operative planning, open-loop control.
- ~15% surgery complication rate.

[Nasser et al. (2010)]

Can RL address these limitations?

- Non-invasive but more noisy & complex observations (Ultrasound, RGBD).
- Closed-loop control, intra-operative planning.



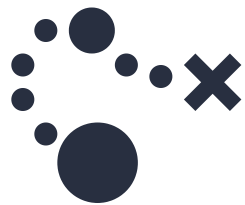
Safety in Reinforcement Learning

Three common approaches

Ergodicity

[Moldoven & Abeel (2012)] [Turchetta et al., (2016)]

[Eysenbach et al., (2017)]



Lyapunov Stability

[Berkenkamp et al., (2017)]

Constrained Markov Decision Processes

[Altman, (1999)] [Achiam et al., (2017)]

[Dalal et al., (2018)]

Safe adaptation via Meta-Learning

How can agents adapt efficiently and safely to new tasks?

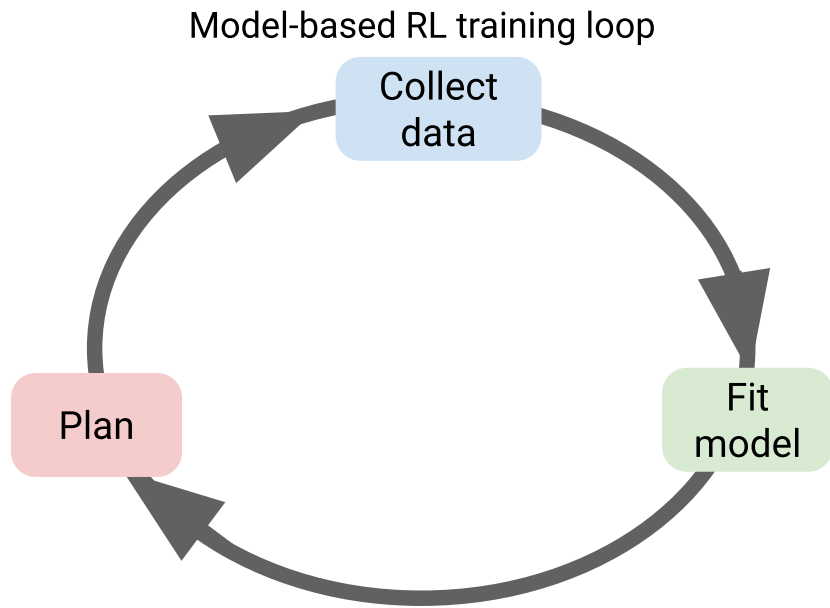
Paper	Safety?
`[Duan et al., (2016)]`, `[Finn et al., (2017)]`, `[Rothfuss et al. (2018)]`, `[Nagabandi et al. (2018)]` and more...	❌
`[Zhang et al., (2020)]`, `[Luo et al., (2021)]`	✅

⚠️ Current literature on *safe* adaptation does not address most of its challenges.

Model-Based Reinforcement Learning

Using supervised-learning to accelerate

- Collect data on the real environment.
- Use this data to fit a statistical model of the environment.
- Use the model to (cheaply) simulate trajectories for policy learning or online control.

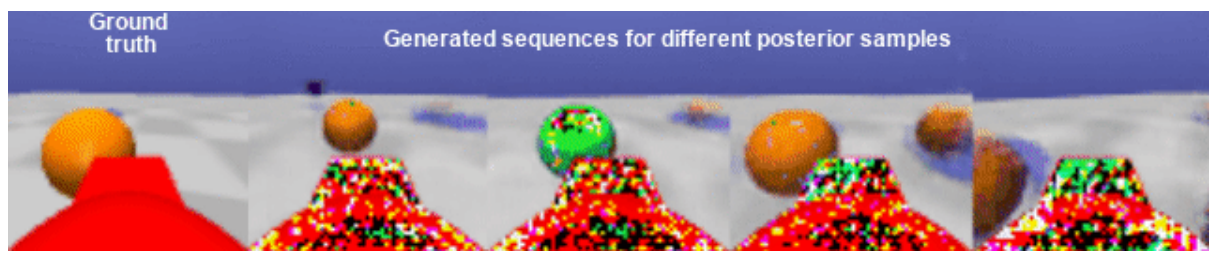


LAMBDA

Constrained Policy Optimization via Bayesian World Models (ICLR 2022), joint work with Ilnura Usmanova, Sebastian Curi and Andreas Krause

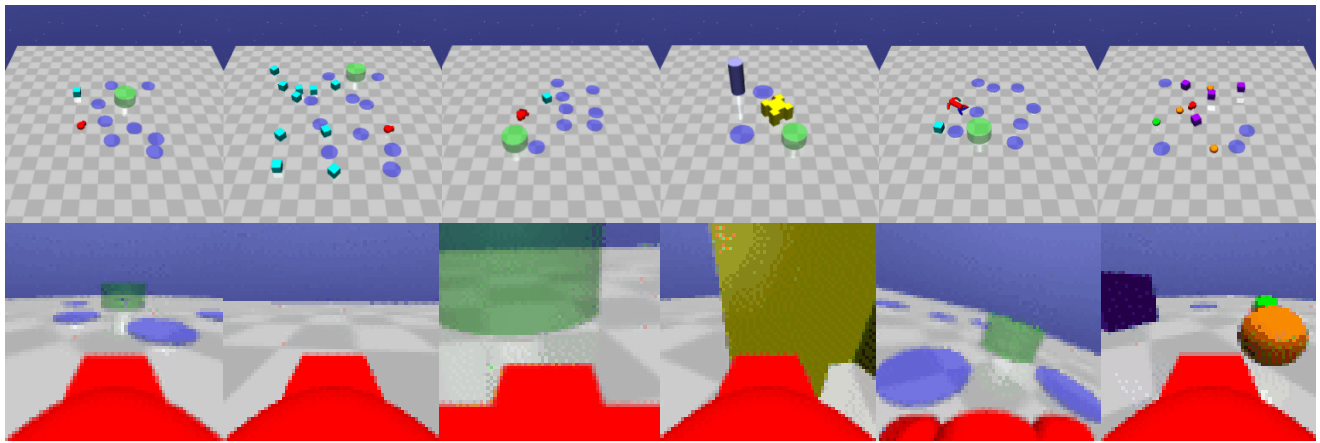
- Maintain a posterior distribution over model parameters given previously seen data $\theta \sim p(\theta|\mathcal{D})$
- Models track an underlying representation of the environment's state, given image observations (think non-linear Kalman filter)

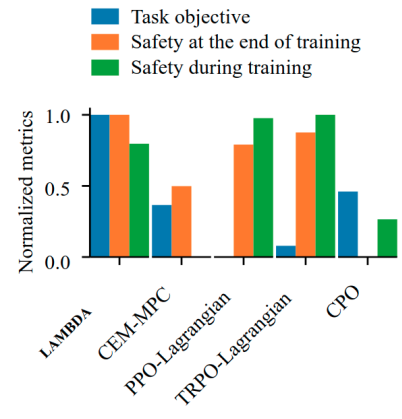
This probabilistic modeling allows the agent to be robust for safety (through pessimism); but still discover new behaviors (through optimism).



LAMBDA

Testing LAMBDA with the Safety Gym benchmark suite





Progress to Date

Creating a testbed for safe adaptation algorithms

`safe-adaptation-gym`

A benchmark suite for safe adaptation

- 8 different tasks.
- Each sampled task is subject to different dynamical properties.
- Number of obtacles and their sizes sampled randomly per task.

`safe-adaptation-agents`

- Implementation of 4 different baseline meta-RL algorithms.
- Use common CMDP solvers for safety.